UNIVERSITY OF CAMBRIDGE

M. PHIL. IN STATISTICAL SCIENCE

Tuesday 12 June 2007 9 to 12

APPLIED STATISTICS

*Attempt **THREE** questions.*

*There are **FOUR** questions in total.*

The questions carry equal weight.

This is an 'Open Book' examination, involving the use of the Statistical Laboratory's network of workstations. Candidates will receive this paper at 9.00am on Tuesday 12 June, and must hand in their scripts to the Chairman of Examiners by 1.00pm on Friday 15 June.

The data sets will be emailed to candidates on Tuesday 12 June.

(The Statistical Laboratory Computer Officer and an Examiner will normally be available for consultation if required between 9.00am and 4.30pm on these four days.)

Each candidate should submit his/her script with a signed statement that the work has been carried out without any collaboration with others.

The scripts may be handwritten. Candidates are requested to submit at most 25 pages in total. They are advised that the total work set should take between 4 and 6 hours.

STATIONERY REQUIREMENTS
Cover sheet
Treasury Tag
Script paper

SPECIAL REQUIREMENTS
None

You may not start to read the questions printed on the subsequent pages until instructed to do so by the Invigilator.

1 The Times, 5 June 2006, published their Good University Guide giving data on nine aspects of activity for 109 universities. The nine aspects are

Satis	a score of student satisfaction with teaching taken from a national student survey (maximum possible score is 20)
Res	average quality of research per staff member (maximum possible score is 7)
Entry	average A level entry score
Stud.staff	student:staff ratio
Lib	library/computer spending in £ per student
Facil	facilities spending in £ per student
GoodHons	percentage of graduates with a First or Upper Second Class Honours degree
Prosp	percentage of graduates in "graduate" employment or further study
Complet	percentage of students who graduate

A few lines of the data are given below. The final column is the overall score which is calculated form the above nine aspects of activity

Rank	Inst	Satis	Res	Entry	Stud:staff	Lib	Facil	GoodHons	Prosp	Complet	Total
1	Oxford	NA	6.5	511.7	13	1656	364	88.4	74.8	97.7	1000
2	Cambridge	NA	6.6	525.1	11.9	1129	425	84.6	86.9	98.9	973
3	ImperialCollege	14.4	6.4	468.2	9.4	1230	481	75	83.8	96	878
.											
.											
.											
108	Luton	14.1	1.8	185.7	23.2	606	229	48.4	41.7	74.8	302
109	ThamesValley	NA	0.5	202.5	23.7	398	57	48.8	60.1	68.6	281

(i) Summarise the data using appropriate plots and tables.

(ii) Investigate how the proportion of graduates with a First or Upper Second Class Honours degree depends on the other eight aspects of activity, carefully explaining any model selection techniques that you use in your analysis and illustrating the use of `boxcox( )`. Summarise your conclusions.

(iii) Discuss briefly the choice of the nine variables included in the overall score for the above ranking of universities. What other variables might have been included?

2 The data below are taken from a blood pressure study of insured lives (North American Actuarial Journal, 2003). The table shows the actual (A) and expected (E) number of deaths for 299 933 life insurance policies in the US between 1975 and 2001, cross-classified according to the policyholders' systolic and diastolic blood pressure (BP, in mm Hg) at the time of issue. The expected number of deaths in a particular cross-classification is the number of deaths expected for that particular group according to standard mortality tables.

	E	A
Systolic BP 80-125		
Diastolic BP 50-75	2003.1	1504
Diastolic BP 76-80	1326.0	1042
Diastolic BP 81-85	407.7	332
Diastolic BP 86-90	144.6	106
Diastolic BP 91-112	15.5	18
Systolic BP 126-130		
Diastolic BP 50-75	353.0	307
Diastolic BP 76-80	687.2	596
Diastolic BP 81-85	372.2	334
Diastolic BP 86-90	193.6	194
Diastolic BP 91-112	26.2	23
Systolic BP 131-135		
Diastolic BP 50-75	168.0	137
Diastolic BP 76-80	294.9	308
Diastolic BP 81-85	356.5	318
Diastolic BP 86-90	197.0	160
Diastolic BP 91-112	49.7	50
Systolic BP 136-140		
Diastolic BP 50-75	144.6	136
Diastolic BP 76-80	357.5	378
Diastolic BP 81-85	329.8	305
Diastolic BP 86-90	459.0	443
Diastolic BP 91-112	118.6	97
Systolic BP 141-145		
Diastolic BP 50-75	25.5	31
Diastolic BP 76-80	75.6	85
Diastolic BP 81-85	128.5	123
Diastolic BP 86-90	174.9	190
Diastolic BP 91-112	134.1	133
Systolic BP 146-230		
Diastolic BP 50-75	75.4	89
Diastolic BP 76-80	125.4	153

Diastolic BP 81-85	177.3	187
Diastolic BP 86-90	286.0	319
Diastolic BP 91-112	401.6	549

BP=Blood Pressure

The aim is to investigate possible multiplicative effects of systolic and diastolic blood pressure on the mean number of deaths relative to E.

(i) Carry out exploratory plots and summaries of the data.

(ii) Assume A is Poisson with mean μE. Treating systolic and diastolic bloody pressure as factors, fit a model for μ in terms of the levels of systolic and diastolic blood pressure.

(iii) The mortality ratio is defined as A/E. Using your preferred model in (ii), give an estimate and standard error for the expected mortality ratio when systolic and diastolic blood pressures lie in the ranges 126-130 and 91-112 respectively, and comment briefly.

3 The tables below are taken (slightly edited) from the Healthcare Commission annual health check (October 2006). Table 1 shows the Quality of Service ratings (Excellent, Good, Fair, Weak) for Healthcare Trusts in four regions (North, Central, London and South East, South West)

TABLE 1

	Excellent	Good	Fair	Weak	Total number of Trusts
North	10	63	83	14	170
Central	3	74	86	9	172
London & South East	5	39	65	11	120
South West	5	30	54	16	105

(i) A Trust is satisfactory if it is classified as Excellent or Good, and is otherwise unsatisfactory. Investigate whether the proportion of satisfactory Trusts varies over the regions using (a) a χ^2 test and (b) binomial modelling.

(ii) For the original four Quality of Service ratings in Table 1, use Poisson models to investigate whether there are differences in Quality of Service ratings between regions.

(iii) Suppose that further examination of the data revealed that Table 1 was obtained by adding together the results for four different types of Trust. For the North and Central regions these results are shown in the tables below.

Type of Trust		Excellent	Good	Fair	Weak
Acute & Specialist	North	5	31	13	3
	Central	1	19	26	1
Ambulance	North	0	4	0	4
	Central	0	5	1	3
Primary care	North	3	24	59	6
	Central	2	42	50	5
Mental Health	North	2	4	11	1
	Central	0	8	9	0

What do you conclude now about differences between the North and Central regions in the proportion of satisfactory Trusts?

4 Shown below is a subset of a dataset from a study commissioned by an Investment Bank to look at the performance of its traders in its various branches across the financial world. Over a two-week period, information on the proportion of transactions that were "successful" for each trader was collected. In addition, covariate information on age, gender, and years of experience of the trader and the location of the branch where the trader is based was recorded.

id	sex	age	branch	experience	Ntrans	Nsucc
1	1	29.03	Tokyo	3.52	114	87
2	1	36.75	Tokyo	2.61	161	118
3	1	33.87	Tokyo	3.20	101	73
4	1	34.57	Tokyo	2.56	150	111
.						
.						
.						
199	1	25.56	New York	2.85	150	97
200	1	29.88	New York	3.08	119	84

```
id  = Trader's anonymous identifier
sex = Sex of trader (0 corresponds to female; 1 to male)
age = Age of trader (in years)
branch = Branch Location
experience = Years of experience as a trader
Ntrans = Total number of transactions conducted over two-week period
Nsucc = Number of ``successful'' transactions
```

The Consulting Firm, hired by the Investment Bank to look at the data, was set the remit of developing an appropriate model for the proportion of successful transactions completed during the two-week study period, in terms of the covariate information collected. This model should allow the Bank to assess the performance of each branch, and investigate the variables associated with good trader performance.

After lengthy discussions with a panel of Executive Members from the London Branch of the Bank, the Consulting Analyst Team assigned to this project elicited the following opinions from the panel:

(i) The panel expressed the opinion that traders within the early years of their career are more motivated and challenged on average than traders later on in their careers.

(ii) They felt that the performance of male and female traders may vary differentially with age.

(iii) They felt that different "cultural attitudes" amongst the different stock markets may affect traders' performances. The Financial Times London Stock Exchange (FTSE) was considered the most important Stock Exchange for their overall business.

You, as the only statistician on the Consulting Team, have been given the task of analysing the data.

(a) Begin by exploring the data using summary statistics, cross-tabulations and plots. Note anything that is helpful in understanding the data. What are the average success rates overall, by branch, and by gender?

(b) By appropriately modelling the data and with the use of plots, investigate the functional relationships (possibly non-linear) of the variables on the probability of a successful transaction. Include generalised linear models and generalised additive models in your analysis.

(c) Determine whether the opinions of the panel are substantiated by the data collected. Interpret and make inferences on your findings. Note that you should compare and contrast the different models you have fitted and then decide on the "best" model to use to base your final inferences on. You need to justify statistically the choice of "best" model.

(d) Is the proportion of successful transactions made a good measure of trader's performance? What other measures of performance could be used?

END OF PAPER